RMIT and Gunma University at NTCIR-9 GeoTime Task

Michiko Yasukawa*  J Shane Culpepper †
Falk Scholer †    Matthias Petri †

*Gunma University, Japan
† RMIT University, Australia
Table of Contents

- Background
- Experimental Framework
- Results
- Conclusions
Background

- Inverted indexes
  - A classical solution for search problems.
    - A vocabulary of terms mapped to documents.
  - Terms (words or n-grams) are defined at indexing time, and not changed at query time.

- Self-indexes
  - A new viable alternative to inverted indexes.
    - A data structure for character level pattern matching.
    - Word boundaries are flexibly changed at query time.
    - Search terms are arbitrary patterns of characters.
Ranked Self-Indexing

Prior work
- Frequency counting for a single phrase.
- Search effectiveness has not been evaluated.

A new search engine, NeWT [Culpepper, et al. 2010]
- Efficient term frequency counting.
  - two wavelet trees
  - BWT (Burrows-Wheeler Transform)
- Anything can be a term at query time.
  - Ranked search for multiple phrases, words, morphemes, and/or any character sequences.
Ranking metrics in NeWT

* (1) raw term frequency:

\[
\text{RAW} = \sum_{t \in q} f_{t,d}
\]

RAW : the aggregate of the term frequency, \( f_{t,d} \).

\( f_{t,d} \) : term frequency counts per document.

* (2) BM25 variant:

\[
\text{BM25} = \sum_{t \in q} \log\left( \frac{N - f_t + 0.5}{f_t + 0.5} \right) \cdot \text{TFBM25}
\]

\[
\text{TFBM25} = \frac{f_{t,d} \cdot (k_1 + 1)}{f_{t,d} + k_1 \cdot ((1 - b) + (b \cdot l_d / l_{avg}))}
\]

\( N \) : the number of documents in the collection.

\( f_t \) : the number of distinct documents appearances of \( t \).

\( l_d \) : the number of UTF8 symbols in the documents.

\( l_{avg} \) : the average of \( l_d \) over the collection. \( k_1 = 1.2, b = 0.75 \)
Our Goal for NTCIR-9 GeoTime Task

- Compare the search effectiveness:

  Indri $\rightarrow$ *classical*
  - Inverted index
    (Terms are static.)
  - Multilingual support

  NeWT $\rightarrow$ *innovative*
  - Self-index
    (Terms are flexible.)
  - Language independent

- [Step1] Search in English with Indri.
- [Step2] Experiment in English with NeWT.
- [Step4] Experiment in Japanese with NeWT.
- [Step5] Query Expansion in Japanese with NeWT.
Experimental Framework (for English)

- Step 1: English search with Indri
- Step 2: English search with NeWT
Experimental Framework (for Japanese)

- Step 3: Japanese search with Indri
- Step 4: Japanese search with NeWT → Substring Mismatch
- Step 5: Step 4 + $n$-suffix query expansion
Query Expansion in Japanese

- Boolean search to gather initial documents.
  - All topic terms appear in each document. (AND)
  - Synonyms from Japanese WordNet and Wikipedia. (OR)
- Later documents likely contain “when and where”.
  - Reverse chronological order of time stamp. (ORDER BY)

![Diagram showing the flow of query expansion with Japanese WordNet and Wikipedia as inputs, PostgreSQL as the database, and documents containing all primary terms as output.](image)
Query Expansion in Japanese (Cont.)

- Regular Expression in PostgreSQL
  - \( n \)-suffixes from the gathered documents. 
    \( (n \)-character suffixes at the tail of the query term)

- For the experiment:
  - 100 \( n \)-suffixes per topic.
  - \( n \)-suffixes using \( n=2, 3, 4 \).

```
<table>
<thead>
<tr>
<th>Regular Expression</th>
<th>n-suffixes</th>
</tr>
</thead>
<tbody>
<tr>
<td>('peter piper', 'p.{1}', 'g')</td>
<td>{pe}, {pi}, {pe}</td>
</tr>
<tr>
<td>('peter piper', 'p.{2}', 'g')</td>
<td>{pet}, {pip}</td>
</tr>
<tr>
<td>('peter piper', 'p.{3}', 'g')</td>
<td>{pete}, {pipe}</td>
</tr>
<tr>
<td>('peter piper', 'p.{4}', 'g')</td>
<td>{peter}, {piper}</td>
</tr>
<tr>
<td>('peter piper', 'p.{8}', 'g')</td>
<td>{peter pip}</td>
</tr>
</tbody>
</table>
```
Results in English

- NeWT run EN-01 shows higher performance. (nDCG@10)
- But, more poorly on other effectiveness measures.
- Overall, no statistically significant difference.

<table>
<thead>
<tr>
<th>Run</th>
<th>System</th>
<th>Ranking</th>
<th>Preprocess</th>
<th>Expansion</th>
<th>MAP</th>
<th>Q</th>
<th>nDCG@10</th>
<th>@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-01</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>None</td>
<td>0.2477</td>
<td>0.2524</td>
<td>0.4282</td>
<td>0.3691</td>
</tr>
<tr>
<td>EN-02</td>
<td>Indri</td>
<td>Dirichlet LM</td>
<td>Krovetz</td>
<td>None</td>
<td>0.2830</td>
<td>0.3057</td>
<td>0.3531</td>
<td>0.3763</td>
</tr>
<tr>
<td>JA-01</td>
<td>Indri</td>
<td>Dirichlet LM</td>
<td>ChaSen</td>
<td>None</td>
<td>0.3779</td>
<td>0.4119</td>
<td>0.4769</td>
<td>0.5109</td>
</tr>
<tr>
<td>JA-02</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>None</td>
<td>0.3084†</td>
<td>0.3239†</td>
<td>0.3510†</td>
<td>0.3936‡</td>
</tr>
<tr>
<td>JA-03</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>2-suffixes</td>
<td>0.3282</td>
<td>0.3349</td>
<td>0.4768</td>
<td>0.4653</td>
</tr>
<tr>
<td>JA-04</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>3-suffixes</td>
<td>0.3671</td>
<td>0.3714</td>
<td>0.5230</td>
<td>0.5211</td>
</tr>
<tr>
<td>JA-05</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>4-suffixes</td>
<td>0.3376</td>
<td>0.3398</td>
<td>0.4988</td>
<td>0.4841</td>
</tr>
</tbody>
</table>

† and ‡ indicate statistical significance relative to the baseline run at the 0.05 and 0.001 levels respectively, based on a paired t-test.
Results in Japanese

- The NeWT run JA-02 performed worse than the Indri run JA-01.
- The 3- and 4-suffix query expansion runs were effective. (nDCG@10)
- But, the differences were not statistically significant.

<table>
<thead>
<tr>
<th>Run</th>
<th>System</th>
<th>Ranking</th>
<th>Preprocess</th>
<th>Expansion</th>
<th>MAP</th>
<th>Q</th>
<th>nDCG@10</th>
<th>@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-01</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>None</td>
<td>0.2477</td>
<td>0.2524</td>
<td>0.4282</td>
<td>0.3691</td>
</tr>
<tr>
<td>EN-02</td>
<td>Indri</td>
<td>Dirichlet LM</td>
<td>Krovetz</td>
<td>None</td>
<td>0.2830</td>
<td>0.3057</td>
<td>0.3531</td>
<td>0.3763</td>
</tr>
<tr>
<td>JA-01</td>
<td>Indri</td>
<td>Dirichlet LM</td>
<td>ChaSen</td>
<td>None</td>
<td>0.3779</td>
<td>0.4119</td>
<td>0.4769</td>
<td>0.5109</td>
</tr>
<tr>
<td>JA-02</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>None</td>
<td>0.3084‡</td>
<td>0.3239†</td>
<td>0.3510†</td>
<td>0.3936‡</td>
</tr>
<tr>
<td>JA-03</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>2-suffixes</td>
<td>0.3282</td>
<td>0.3349</td>
<td>0.4768</td>
<td>0.4653</td>
</tr>
<tr>
<td>JA-04</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>3-suffixes</td>
<td>0.3671</td>
<td>0.3714</td>
<td>0.5230</td>
<td>0.5211</td>
</tr>
<tr>
<td>JA-05</td>
<td>NeWT</td>
<td>BM25</td>
<td>None</td>
<td>4-suffixes</td>
<td>0.3376</td>
<td>0.3398</td>
<td>0.4988</td>
<td>0.4841</td>
</tr>
</tbody>
</table>

‡ and † indicate statistical significance relative to the baseline run at the 0.05 and 0.001 levels respectively, based on a paired t-test.
Conclusions

- A new self-indexing search engine, NeWT
  - Experimented on the multilingual task.
    - Language processing at query time, not at indexing time. 😄
    - Multiple languages can be incorporated into a single index. 😊
  - Search effectiveness was examined.
    - Efficient document ranking with self-indexes. 😊
    - For GeoTime topics, no significant effectiveness. 😞

- Future work:
  - Efficiently determine IDF (Inverse Document Frequency).
  - Explore the substring mismatch problem.
Thank you very much for your kind attention.

Michiko Yasukawa
michi@cs.gunma-u.ac.jp